文献相似性对科学引用偏好的影响实证研究*

■ 段庆锋[1] 潘小换[2]

[1] 山西财经大学管理科学与工程学院 太原 030006 [2] 中北大学经济与管理学院 太原 030051

摘要:[**目的/意义**]施引文献与被引文献往往存在着某种相似性,揭示这种现象背后的形成机制有助于深入理解引文的本质。[**方法/过程**]采用指数随机图模型,以图书馆与情报学领域为对象开展实证分析,旨在揭示文献相似性对引用关系的影响机制。[**结果/结论**]实证研究发现:在网络结构、机构、期刊层面存在显著的引用文献相似倾向。具体地,引用关系更倾向于嵌入三角传递结构;来源于相同机构和期刊的文献之间更容易产生引用关系;来源于学科优势地位国家的文献之间更容易产生引用。实证结果充分说明社会接近性是引用行为的重要形成机制,反映了引用偏好的社会属性。

关键词:文献 相似性 科学引用 指数随机图模型

分类号:G253

DOI:10.13266/j.issn.0252-3116.2018.04.013

1 引言

在现代科学发展中,规范化的文献引用扮演了重要角色,也激发了学者们研究其内涵及机理的兴趣。尤其值得关注的是引文已经成为现代学术评价的重要理论基石和度量工具,以期刊影响因子、h指数为代表的评价工具都离不开引文。不可忽视的问题是:建立在引文基础上的评价理论和工具都实质上隐含了一个关键而理想化的前提,即引用对象的选择是完全基于学术价值判断或某种学术目的的。显然,严格的前提假设与实际应用之间存在较大差距,实践中出现的各种争议和学界的争论都反映了引文内涵的复杂性[1]。因此,非常有必要追溯到问题的本源,探究引用形成的机制,对引文价值的准确理解可能有助于提出或修正更具学术价值甄别效能的引文评价指标。

考虑到引用本质上是文献间的二元关系,因此施引文献与被引文献之间的相似性可能是考察引用内在机制的有效视角和途径。这种相似性可能表现出多种不同形式,例如,在文献作者层面的相似特征。如学术界早已发现学科中常存在某个高频率相互引用的核心圈子,活跃在其中的作者往往都是领域中的高影响力学者[2]。引用关系聚集在特定核心学者群体中并不是偶然现象,这种所谓的学术圈子正是某种社会接近性的体现。有学者提出"精英俱乐部"指数[3],以引用关系为线索,从统计的角度识别出这些相互高度认同的学者群体。

自引可以被视为一种独特引用现象,即施引文献与被引文献来源于完全相同的作者。针对自引现象的研究非常丰富,尤其聚焦于它对科学评价产生的公平性和适用性问题[4]。一方面,有学者认为不受控制的自引数量会扭曲合理正常的引文分布,包含自引的评价指标的公正性受到质疑[5];另一方面,也有学者认为针对自引情形需要具体分析,当样本足够大时没有必要剔除自引,甚至有问卷调研发现科研人员在进行自引和他引时并不存在显著差异的动机[6]。

伴随引用关系出现的文献相似性倾向可能表现在多个层面,除了文献作者之外,国家或地区、语言、期刊、机构可能都存在类似的引用偏好倾向。例如A. Bookstein和M. Yitzhaki[7]设计了母语偏好指标来研究相同语言群体间的引用倾向,说明了语言偏好性对引用行为的影响。S. Ren和R. Rousseau[8]从期刊引用的角度开展实证研究,发现中国期刊之间的高比例互引现象;唐莉等[9]分析了中国科研成果高速增长背后

* 本文系山西省高等学校哲学社会科学研究基地项目"山西省煤基低碳产业创新链技术预测研究"(项目编号:2016325)和山西省软科学项目"山西省战略新兴产业技术创新潜力及研发方向研究"(项目编号:2017041003-1)研究成果之一。

作者简介:段庆锋(ORCID:0000-0002-8008-5563),副教授,博士,E-mail:dqf01@sina.com;潘小换(ORCID:0000-0002-8970-7250),硕士研究生。

收稿日期:2017-08-17 修回日期:2017-11-28 本文起止页码:97-106 本文责任编辑:易飞

存在的"俱乐部效应",通过中美对比分析说明高被引论文中中国作者内部存在更为显著的互引行为。类似现象的广泛存在很大程度上暗示了社会属性对于引用偏好的重要影响性,因此从更广义的社会接近性视角探讨施引文献和被引文献在某些属性上表现出的相似或接近倾向,能有助于更系统地揭示引用偏好的形成机制。

引用关系中出现的文献趋同现象虽然已引起部分学者关注,但研究文献较为零散而不够系统,而且学术界对其背后的机制探讨与理论分析不够深入[10]。通过梳理相关文献,发现这些研究存在以下不足:①理论探讨较多,实证分析匮乏,尤其是缺乏充分利用文献大数据的建模分析;②引用过程中文献趋同性是具有多维表现的,但已有文献大多只是聚焦于单个维度,缺乏统一分析框架下的系统性讨论;③大多采用描述性分析,缺乏统计推断的分析,难以判定文献相似性到底在多大程度上影响了引用偏好的形成。

由此,引出本文所关注的基本问题:伴随引用关系的文献相似性普遍存在吗?它们在多大程度上影响了引用关系的形成?通过实证结果分析,探究外部社会因素在学术引用过程中的影响,对这些问题的深入思考有助于完善对于科学引文本质的理解。

因此,以社会建构理论为指导,提出不同层面下引用行为中文献相似性倾向的研究假设;以图书馆与情报学领域为实证学科领域,采用指数随机图模型,以引用关系为因变量,将引用关系形成的概率建模为有关文献相似性指标变量;最终通过实证结果分析,探讨引用形成的基本机制。

2 研究方法

2.1 指数随机图模型

随机指数图模型(ERGM)是一种针对网络二元关系构建的计量分析模型,也被称为 P^* 模型,已经成为一种重要而应用广泛的针对网络边的建模方法,可以用来刻画不同因素对网络二元关系形成的影响。模型将观察到事实网络 y 的概率建模为各种可能构型(configuration),例如边数、三角形、互惠性等,以及节点属性和边属性。ERGM 模型定义[11]的形式如公式(1)所示:

$$p(Y=y|X)=\frac{\exp\{\theta^T g(y,X)\}}{x(\theta,y)} \quad \text{公式(1)}$$

其中 Y 是随机关系的集合,可以用随机邻接矩阵 Y_{ij} 表示,第 i 行 j 列的元素对应于从第 i 节点到第 j 节点的关系;y 是随机邻接矩阵 Y 的一个实现,是特定的观察到的关系。X 是与边或节点有关的协变量组成的向量。θ 是系数向量,是各种变量前的相应系数。$g(y, X)$ 是由网络变量构成的向量,如果某构型在网络 y 中被观察到 k 次,则 $g(y)=k$。$x(\theta,y)$ 是归一化因子,以确保所有可能网络样本出现的概率和为1,即 $\sum_{z\in Y}\exp\{\theta^T g(z,X)\}=1$。

为了进一步说明参数 θ 的解释能力,引入变化统计量 δ,如公式(2)所示:

$$\delta_g(y)_{ij}=g(y_{ij}^+-g(y_{ij}^-) \quad \text{公式(2)}$$

其中 y_{ij} 代表节点 i 与 j 的二元关系,1 代表 i 与 j 存在连接关系,否则为 0。y_{ij}^+ 和 y_{ij}^1 分别代表在 y 的其余部分保持不变的情形下,分别设定 $y_{ij}=1$ 或 $y_{ij}=0$ 的网络实现。也就是说 $\delta_g(y)_{ij}$ 反映了当 y_{ij} 由 0 变为 1 而且其他边保持不变的条件下,网络统计量 $g(y)$ 的变化量。利用变化统计量,可以将公式(1)等价地转化为另一种形式,如公式(3)所示:

$$\text{logit}[P_{\theta,y}(Y_{ij}=1|Y_{ij}^c=y_{ij}^c)]=\theta^T\delta_g(y)_{ij} \quad \text{公式(3)}$$

其中函数 logit 代表对数几率,定义为 $\text{logit}(p)=\log[p/(1-p)]$,$P$ 为二元关系发生的概率,Y_{ij}^c 代表除了网络边 Y_{ij} 之外的网络 Y 的其余部分。公式左边代表了网络其余部分 Y_{ij}^c 不变的情形下节点 i 和 j 建立连接的对数几率。公式右边代表了,固定网络其余部分不变,当 y_{ij} 由 0 变为 1 时,网络统计量 $g(y_{ij})$ 每增加 1 个单位,则 i 和 j 建立连接的概率是不建立连接概率的 $exp(\theta)$ 倍。参数 θ 的大小反映了各种因素对网络边关系的边际效应。

采用指数随机图模型进行建模分析,主要考虑以下两点:①文献通过引用关系相互连接,形成引用网络,专门针对网络二元关系建模的指数随机图模型非常适用于网络样本;②指数随机图模型比传统回归模型更适合于网络关系建模,充分考虑了网络边关系存在的自相关性,满足了网络建模的要求;③指数随机图模型能够将引用二元关系发生的概率解释为引用网络内生构型和外生协变量的函数,将各种相似性变量加入模型,其统计推断功能有助于识别出不同形式文献相似性对引用关系的影响程度。

2.2 研究假设

从引用网络形成的角度看,影响每个引用关系的因素可以大致划分为两类:①各个引用关系存在着相互的影响,由于网络边自相关性带来的内生性结构因素是需要考虑的内容;②源于网络节点或边的外生性

属性也需要关注的要素。

2.2.1 内生性结构趋同倾向 引用关系内嵌于文献网络之中,从网络视角开展探讨有助于更深入地揭示引用行为的复杂性与多元特征。网络系统中各条边并非独立存在的,而是相互依赖与影响,因而在拓扑上可能出现某种稳定的结构。从网络形成的机制角度看,网络边建立过程可能受到这种内生性影响而产生某种连接倾向,以更大的概率形成某种连接关系以获取某种网络优势,嵌入结构模式之中的节点也受益于网络效应。结构特征是社会网络分析所关注的方面,它们可能体现在网络系统的不同尺度层面,例如宏观层面的聚类、中观尺度的社区,而在微观层面上最为简单而基本的结构便是三角结构。简单地说,以经典的朋友网络为例,所谓的"朋友的朋友可能也是朋友"则正反映了这种基本三元结构,在期刊引证网络[12]、合作网络[13]为代表的实际网络中都证实了这种连接模式的存在性。

引用网络是有向网络,传递三角结构是其中一种基本结构。一般地,如果文献 i 引用文献 j,文献 j 又引用了 k,则文献 i 也会引用文献 k。嵌入传递三角结构之中,文献 i 和文献 k 之间存在着冗余路径,除了直接连接还有 1 条长度为 2 的间接路径连接。从知识流的角度,文献 j 在文献 i 和 k 之间扮演着知识中介者的角色,它吸收一方的知识,并将增值后的知识再传递给另一方,第三方文献共享机制为两文献提供了潜在的知识转移渠道,避免了新建知识流动渠道所增加的成本和风险。因此,非直接连接的文献共享相同的知识,它们之间存在建立直接连接关系的倾向,形成传递三角结构。从知识流动的效率看,冗余路径结构提供了更为高效的知识传播体系,降低了由于系统结构层面的脆弱性导致的知识传递链断裂风险,具有更高的知识传递网络鲁棒性。传递三角结构嵌入性反映科学文献在引用网络微观层面的结构趋同倾向。因此,在社会网络理论的基础上提出研究假设:

假设 1:科学引用关系倾向于嵌入传递三角结构。

2.2.2 外生性趋同倾向 科学引用形成机制的指导理论中,社会建构理论获得了广泛的关注。它认为引用行为的产生更多地是一个社会的过程,受到了外部政治、经济等社会要素的多重作用和影响,产生引用关系的驱动因素并不局限于纯粹的学术范畴,而是具有更为复杂和多元的社会属性。例如,L. Bornmann 和 H. Daniel 系统梳理关于引用动机的研究,着重强调了非学术因素在科学引用生成中的影响[14]。该理论为施引文献与被引文献表现出的非学术关联现象提供了解释,从更为广泛的社会关系开展探讨有助于理解引文的复杂性。

如果说施引文献与被引文献之间的某种社会关系不是偶然出现的,而是与引用关系共生,则它们两者之间可能存在着某种不可忽视的内在联系,如果能够准确地揭示两者之间的依赖,必然有助于更加全面地认知引用的内在本质。从知识流动的角度看,新的知识或观点更容易流向吸收或接收能力强的文献节点,嵌入相近知识社会网络的观点更容易被接纳。两篇文献如果在各个社会属性方面越接近,则它们更容易建立显性的引用关系。

如果将科学文献视为知识生产的最终结果,则知识生产系统中包含与涉及的有关要素都是不可缺少的,创新主体是知识的生产、传播与吸收者,知识传播离不开有形的承载客体,知识流动与溢出受到空间及组织的边界约束。依据社会建构理论,引文不但是无形知识相互影响和碰撞的体现,更是环境要素交织作用的结果;除了心理、智力、知识因素,引用偏好形成与知识流动方向还应该是社会因素的结果。因此,科学文献相似倾向可能表现在以下多个方面:

(1)文献载体的同配。作为学术知识载体,每个期刊往往体现出了鲜明的学科领域、选题内容方面的个性标准与偏好,意味着相同期刊上的文献具有更相似的知识结构,而刊载于不同期刊的文献则相反。引文根本上是知识交流与衍生的外在表象[15],而同质化的知识更容易被吸收和理解。当然除此之外,也可能存在由于追求期刊影响因子而导致的出版商和投稿者间形成的不当过度自引,虽然这样的情形可能是个案,但也不可忽视。基于此,本研究提出假设:

假设 2:相同期刊的文献更容易发生引用关系。

(2)正式组织的同配。科学研究早已成为职业,科研人员隶属于某个学术单位,这些单位内部的创新主体自然形成了长期而稳定的学术关系。相同的学术单位意味着学术关系的嵌入,这种相互依赖性体现了体制、制度、组织层面的安排与保障。社会网络强关系不但是显性知识传递的路径,长期合作与面对面交流更为隐性知识溢出提供了渠道。基于此,本研究提出假设:

假设 3:源于相同作者单位的文献更容易发生引用关系。

(3)非正式组织的同配。当今科学研究协作与交流趋势日益明显,互联网与社交媒体的发达更是促进

跨组织、跨国家的科学发展,“无形学院”开始成为与正式学术组织互补的科研新模式。学者们形成的群体、社区、圈子是一种松散的耦合创新系统,共同或者相似的领域、兴趣、任务甚至目标是各种非正式组织形成的基础[16]。创新主体自发形成的非正式组织及群体已经成为学者们获取新知识、拓展社会关系的重要渠道,外部社会网络的连接与嵌入可能会为学者带来额外的异质知识、学术资源、声望与优势。基于此,本研究提出假设:

假设 4:作者源于相同非正式组织的文献更容易发生引用关系。

(4)地理空间的同配。新经济地理学对于经济资源的集聚现象开展了深入研究,而作为内生发展动力的科技创新也呈现类似的特征。学术界普遍认为创新主体在地理空间的集聚有利于知识的溢出,这种外部效应促进了创新活动的效率和效果[17]。除了知识溢出的观点之外,可能还隐含着其他社会性因素的可能解释,例如接近的地理空间往往意味着创新主体在本地社会环境的嵌入,包括国家政治、法律制度、语言文化等。这些同质化的外部社会结构为创新主体间的认知与认同提供了基础,提高了学术交流的效果。基于此,本研究提出假设:

假设 5:来源于相近地理空间的文献更容易发生引用关系。

2.3　模型设定与变量选取

表 1 给出了本文模型中需要用到的有关构型变量内涵及其统计量定义,下面结合具体变量加以说明。

表 1　指数随机数图模型网络统计量含义

变量	含义	图例	统计量	假设
Edges	边数		$\sum_{i,j} y_{ij}$	模型常数项,等价于网络密度
Ttriple	传递三角结构		$\sum_{i,k,j} y_{ik} y_{kj} y_{ij}$	引用关系是否倾向于形成闭合模式?
Nodematch(δ)	节点同质性		$\sum_{i,j} y_{ij}\delta_i\delta_j$	是否具有相同 δ 属性的文献更倾向于发生引用关系?
Nodeicov(δ)	节点协变量		$\sum_{i,j} y_{ij}\delta_j$	是否 δ 属性强的文献更倾向于被引用?

(1)内生性结构趋同变量。针对有向引用网络,传递三角结构作为一种网络构型体现了引用关系间的相互依赖性。变量 *Ttriple* 定义为引用网络中传递三角结构的个数,如表 1 所示。如果变量 *Ttriple* 前系数显著为正,则说明相比其他随机结构,引用关系更倾向于嵌入传递三角结构之中,数值越大这种倾向性越大,否则相反。

(2)外生性趋同变量。本文探讨 4 方面的外生性趋同倾向,即文献载体的同配、正式组织的同配、非正式组织的同配、地理空间的同配。如表 1 中所示,如果节点 i 和节点 j 具有相同的分类属性 δ,则记为 1,否则为 0;那么 *Nodematch*(δ)统计量代表了网络中相同属性 δ 的二元组(i, j)数量,即反映了网络中节点在属性 δ 上的匹配程度。

文献载体的同配关系定义为施引文献与被引文献是否发表在相同的期刊。变量 *Nodematch*(*JO*)定义为来源于相同期刊的所有施引文献与被引文献组合的数量。该变量前的系数如果显著为正,说明来源于相同期刊的文献更倾向于产生引用关系,系数数值越大,这种边际效应越强烈,否则相反。

正式组织的同配通过科学文献作者所属机构的匹配关系加以刻画。变量 *Nodematch*(*JO*)定义为具有相同作者机构的所有施引文献与被引文献组合的数量。考虑到科学文献可能存在一篇文献有多名作者而每名作者归属于多个机构的情形,两篇文献只要至少拥有 1 个相同的作者机构,就认为归属机构相同。该变量前的系数如果显著为正,说明来源于相同作者机构的文献更倾向于产生引用关系,系数数值越大,这种边际效应越强烈,否则相反。

非正式组织的同配可以通过科学文献作者是否同样成为高被引作者加以刻画。非正式组织具有多种形式,这里界定为学科领域的高被引作者,这个群体通常在学科领域中具有高的学术影响力和声望,是学科前沿的引领者。具体地,本文将学科领域中累积被引数前 1% 的作者界定为高被引作者。变量 *nodematch*(*AU*)定义为施引文献与被引文献组合的数量,这些文献都是由高被引作者所署名。该变量前的系数如果显著为正,说明引用关系更倾向于出现在高被引作者的文献之间,系数数值越大,这种边际效应越强烈,否则相反。

地理的同配通过文献来源城市的匹配关系加以刻画。本文采用是否同属相同城市来间接作为地理同配关系的代理变量,虽然不如城市间测地线距离精确,但简单可行,同样可以满足研究需要。考虑到科学文献可能会有多个城市地址,如果 2 篇文献至少拥有 1 个相同的来源城市,则认定为来源于相同城市。变量 *Nodematch*(*AU*)定义为具有相同来源城市的施引文献与被引文献组合的数量。该变量前的系数如果显著为正,说明引用关系更倾向于出现在相同来源城市的文献之间,系数数值越大,这种边际效应越强烈,否则相反。

(3)控制变量。模型加入了代表文献学术水平的变量 - 学术价值 *Nodeicov*(*CT*),用于解释学术价值在引用过程中产生的作用效果。本质上,引文是思想与观点的交互与碰撞,文献的思想内容、学术创新性对引用的影响是至关重要的。变量 *Nodeicov*(*CT*)的含义如表 1 所示,定义为被引文献的累积被引数,其中 *CT* 代表文献的累积被引数。该变量用于检验引用网络中被引用文献属性 *CT* 强弱对于连接建立概率的影响。

变量 *Edges* 是事实引用网络中所有边的数量,反映了引用发生的总量。该项属于模型的必选,相当于传统回归模型中的常数项,解释能力等价于网络密度,该变量前的估计系数反映了网络密度对边连接几率影响的边际效应。

模型中引入变量 - 几何加权二元关系共享组合 *GWDSP*(*geometrically weighted dyad-wise shared partners*),其定义为所有可能共享节点二元项分布的加权线性组合。一方面变量表征了网络中形成开放三角结构的倾向,另一方面也有助于降低模型发生退化的风险[18]。

3 实证分析

3.1 数据来源及处理

以图书馆与情报学(LIS)为研究学科领域,选取学科中具有代表性的 7 本国际性期刊为检索范围,包括:*Journal of documentation*、*Scientometrics*、*Journal of information science*、*Electronic library*、*Information technology and libraries*、*Library & information research*、*Journal of the American Society for Information Science*。所有数据来源于 WoS 数据库,样本时间跨度为 1980 - 2010 年,文献类型筛选为 Article,得到初始检索数据 6 111 条。为了便于分析,进一步删去引用网络中的孤立文献。具体地,假设 d 是文献集合 D 中的任意一篇文献,$c \in C$,其中 C 是文献 d 的所有前向引用(forward citation)与后向引用(backward citation)构成的集合,如果 $\forall c \notin D$,则将文献 d 界定为引用网络的孤立文献。依据上述方法,筛选数据,最终得到 2 125 条文献记录数据。

模型分析建立在网络关系基础上,需要从文献元数据中抽取引用关系并形成矩阵数据。将科学文献视为网络节点,如果文献 i 引用文献 j,则形成 1 条由 i 到 j 的有向边。网络矩阵由二元数值构成,设定 1 代表存在引用关系,否则为 0,最终形成 2 125 × 2 125 的引用关系矩阵。同时还抽取了每篇论文的其他有关元数据,包括来源期刊、发表时间、被引数、作者、机构、国家,用于模型中外生协变量的生成。

3.2 描述性分析

对国家、机构、期刊不同层面的描述性分析,有助于初步判断引用文献相似现象的分布规律。这里采用内部引用率来刻画引用趋同程度,指标定义为:源于相同实体(例如国家、机构、期刊)的文献相互引用数量占其施引总量的比例,反映了引用关系中文献的相似程度。如果实体内部引用率越大,说明越倾向于选择源于相同实体的文献作为引用对象。另外,为了识别实体在学科中的影响力,选取被引量为指标,如果实体文献积累的被引数越多,则说明该实体具有越大的学术影响力。表 2 分别给出了按国家和机构统计的内部引用率,列出了被引总量前 10 名的国家和机构情况。

在国家方面,美国的表现无疑一枝独秀,内部引用率达到极高的 70%,反映了美国在该领域处于绝对的领先地位,它们的研究基本能够代表领域前沿水平,从知识流动的角度看形成了内部循环发展的学科生态。影响力前 5 名的其他国家也表现出了较高的内部引用率,分布在区间(32%,39%);其他影响力前 15 名的国家(除了巴西),内部引用率都保持在(10%,20%)之间,其中中国的文献具有 19% 的内部引用率,与整体分布规律比较吻合;值得注意的是巴西 41% 的内部引用率明显高出了其他同水平国家的内部引用倾向,与其影响力水平不相符合的异常知识流动结构可能反映了该国研究较为封闭的特征。

机构方面,被引总量前 15 名的机构大多拥有较高的自引率,分布在区间(20%,50%)之间。以 Leiden Univ、Drexel Univ 等为代表的高校或研究所在图书馆与情报学领域拥有雄厚的实力,内部引用率与被引量两个指标都相对较高。值得注意的是 Inst Sci Informat 的内部引用率高达 55%,追逐该机构的论文可以发现,大部分内部引用论文与科学计量学的奠基人物 H.

表 2 按国家和机构统计的内部引用率(被引总量前 10 名的国家和机构)

序号	国家	内部引用率	被引总量(次)	机构	内部引用率	被引总量(次)
1	USA	0.70	1 260	Leiden Univ	0.30	142
2	England	0.38	380	Drexel Univ	0.47	126
3	Netherlands	0.32	297	Katholieke Univ Leuven	0.35	109
4	Belgium	0.37	238	NatlInst Sci Technol & Dev Studies	0.39	92
5	Hungary	0.39	190	Indiana Univ	0.24	90
6	India	0.14	111	Hungarian AcadSci	0.25	85
7	Canada	0.14	97	Inst SciInformat	0.55	79
8	Spain	0.14	90	City Univ London	0.35	71
9	France	0.17	86	Wolverhampton Univ	0.31	66
10	China	0.19	81	Hungarian AcadSci Lib	0.19	58
11	Denmark	0.18	75	Univ Sheffield	0.28	52
12	Sweden	0.17	65	Limburgs Univ Ctr	0.22	49
13	Finland	0.11	60	Univ N Carolina	0.30	45
14	Brazil	0.41	55	Univ Amsterdam	0.20	43
15	Germany	0.14	50	Royal Sch Lib & Informat Sci	0.03	43

Small 和 E. Garfield 有关,他们所做的开创性研究大多成为当前研究的思想源泉与理论基石。同样,许多具有高内部引用率的机构都可以列出几位具有高影响力的代表性学者,例如 Drexel Univ 有知名学者 K. W. McCain、H. D. White、K. W. McCain、B. C. Griffith,而 Leiden Univ 有知名学者 H. F. Moed、R. J. W. Tijssen、A. F. J. V. VanRaan。科研机构内部的稳定学者群体具有高效率的知识共享和流动便利,也有利于形成高质量水平的科研团队,也是导致科研机构出现较高的内部引用倾向的原因所在。

由表 2 的数据分布可以发现:不论是国家还是机构层面,通过被引数刻画的实体学术影响力与其内部引用率之间似乎存在某种相关性。为了进一步检验这种关联性,表 3 列出了按被引数降序排列的国家和机构内部引用率分布。例如,被引量前 1% 的国家和机构的平均内部引用率分别为 70%、34%;被引量前 50% 的国家和机构的平均内部引用率分别为 34%、18%;所有国家和机构的平均内部引用率分别为 32%、15%。随着学术影响力下降,内部引用率亦呈现梯度下降趋势。

表 3 国家和机构内部引用率分布

被引用总量累积分布	内部引用率均值	
	国家	机构
前 1%	0.70	0.34
前 10%	0.49	0.23
前 25%	0.37	0.20
前 50%	0.34	0.18
前 75%	0.33	0.16
前 100%	0.32	0.15

另外,在国家层面,内部被引率与被引数量呈现正相关,Spearman 秩相关系数为 0.785,通过 1% 显著水平的双侧检验;在机构层面,内部被引率与被引数量呈现正相关,Spearman 秩相关系数为 0.493,通过 1% 显著水平的双侧检验。可以看出,内部被引率与被引数存在一定程度的正相关性,高影响力的国家和机构通常表现出较高内部被引率。

表 4 给出了期刊层面的内部引用率分布。总体上,不同期刊内部引用率指标存在较大的差异。*Scientometrics*、*Journal of the American Society for Information Science* 具有最高的内部引用率,而且不论是被引数还是影响因子都表现出明显优势,反映了它们在学科中具有较高影响力。按照复杂网络理论中"优先连接"机制解释,与其他期刊相比,这两个期刊具有高的影响因子,具有更大的优势与几率被其他文献引用。其他期刊则具有较低的内部引用率,不论是影响因子还是被引数都反映了它们在学术影响力方面稍逊一筹。

3.3 ERGM 模型分析结果

模型参数 θ 的大小与显著性程度是分析各种构型变量对二元因变量影响程度的依据。采用 R 环境中的 STATNET 程序包进行参数估计,具体采用马尔可夫链蒙特卡罗极大似然估计法(MCMC MLE)对模型参数进行检验估计。为了判断参数拟合的效果,使用 t 统计量进行参数显著性的检验。另外,AIC 和 BIC 指标可用于整体模型拟合效果的判断依据。

表 4　按期刊统计的内部引用率

序号	期刊	内部引用率	被引数	影响因子
1	SCIENTOMETRICS	0.82	2015	2.147
2	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE	0.74	946	2.322
3	JOURNAL OF DOCUMENTATION	0.34	297	0.853
4	JOURNAL OF INFORMATION SCIENCE	0.26	257	1.372
5	INFORMATION TECHNOLOGY AND LIBRARIES	0.68	72	1.029
6	LIBRARY & INFORMATION SCIENCE RESEARCH	0.23	71	1.185
7	ELECTRONIC LIBRARY	0.45	53	0.484

注:影响因子来源于 2016 年的 JCR 报告

采用逐步加入变量的策略对模型进行设定和选择,ERGM 模型参数估计结果如表 5 所示。模型 1 只加入了内生性影响因素,而模型 2 在模型 1 的基础上又加入了外生性影响因素。模型 2 中同配变量 *Nodematch*(*AU*)没有通过显著性检验,故将该项删去,最终形成模型 3。与其他模型相比,模型 3 具有最小的 AIC 和 BIC 值,所有统计量参数都通过了 0.1% 水平的显著性检验,说明模型 3 的形式设定是合适的,参数拟合结果也是满意的。下面针对模型 3,进行分析和解释。

表 5　ERGM 模型参数估计结果

变量及指标	模型 1	模型 2	模型 3
内生性结构趋同			
传递三角结构 *Ttriple*	2.076(0.001)***	1.892(0.001)***	1.947(0.000)***
外生性趋同			
期刊 *Nodematch*(*JO*)		1.314(0.055)***	1.285(0.062)***
机构 *Nodematch*(*JG*)		3.440(0.052)***	3.537(0.157)***
高被引作者 *Nodematch*(*AU*)		－0.005(0.031)	
国家 *Nodematch*(*CY*)		－1.123(0.068)***	－0.852(0.094)***
控制变量			
边数 *Edges*	－7.298(0.053)***	－7.692(0.034)***	－7.763(0.074)***
被引数 *Nodeicov*(*CT*)	0.012(0.001)***	0.006(0.001)***	0.064(0.094)***
几何加权二元共享组合 *GWDSP*	－0.131(0.018)***	－0.115(0.010)***	－0.167(0.015)***
AIC	55 675	53 341	51 926
BIC	55 729	53 347	52 020

注:括号中为参数估计量的标准差;***、**、*分别代表 $p<0.001$、$p<0.01$、$p<0.05$

变量 *Ttriple* 的系数显著为正,说明在网络其他部分保持不变条件下,引用关系嵌入传递三角结构中的几率是其他情形的 7($=e^{1.947}$)倍。如果两文献存在间接的引用路径,则它们之间更倾向于建立直接引用关系,体现了在引用网络结构层面的趋同倾向。进一步,正向效应也体现了引用关系形成闭合三角的倾向,结构中各条边之间存在相互条件依赖性,每个引用关系是嵌入网络环境的。基于此,实证结果支持假设 1 成立。

变量 *Nodematch*(*JO*)的系数显著为正,说明了期刊和机构的同配对于建立引用关系具有促进作用。假定引用网络其他条件不变,源于相同期刊的文献发生引用关系的几率是不同期刊情形的 3.61($=e^{1.285}$)倍。一方面,相似的研究领域及选题有利于观点与知识的交流,而相同期刊载体的高可见性也增加了被引的机会。邱均平等[19]采用调查问卷的方法,探究了科研人员论文引用的 5 种动机心理结构与因素,发现信息源便利性对于引用动机的重要作用,而相同期刊来源带来的高可见性很大程度上提高了信息源便利性。另一方面,按照社会建构理论,在出版过程中作为利益相关方的投稿作者与出版方之间存在利益交换的道德风险,这种潜在的引用暗示或默契也有可能推高期刊内部引用比例。基于此,实证结果支持假设 2 成立。

变量 *Nodematch*(*JG*)的系数显著为正,说明了机构的同配对于建立引用关系具有促进作用。假定引用网络其他条件不变,如果两篇文献具有相同的机构,则这两篇文献建立引用关系的几率分别是不同机构情形的 34.36($=e^{3.537}$)倍。值得注意的是:机构同配对于

引用关系建立的影响程度非常大,作用方向虽然与预期一致,但机构同配的影响效应超出了异配情形的34倍。如此强烈的影响效果,反映了社会强关系在科学活动中的重要性。其一,从社会网络的角度看,嵌入相同机构内部社会网络的作者间存在着强关系,这种机构内部的学术合作与交流具有天然的稳定性与低成本,而出现在学术论文中的引用关系指向偏好是同机构作者非正式交流的显性体现;其二,工作在相同机构的作者们通常具有相同或相近的学科背景、知识结构、制度文化、心理距离等,这些可能的相似性都有利于引用关系产生。总之,模型拟合结果很好地支持了正式组织同配的正向效应,因此假设3成立。

变量 *Nodematch*(*AU*)的系数为负,但没有通过显著性检验,说明在高影响力作者内部并不存在更显著的相互引用倾向。这一结论与部分文献所支持的"引用俱乐部"[20]观点不同,原因可能在于:模型所推断检验的是样本整体平均趋势,而学术精英圈子内高频率相互引用的现象只是学科领域整体中的部分,局部的特征无法通过本文模型得到检验;需要结合实际学科领域、针对特定群体加以具体分析。基于此,实证结果不能支持假设4的成立。

变量 *Nodematch*(*CY*)系数显著为负,说明国家层面的文献异配促进了引用关系的建立。具体地,假定引用网络其他条件不变,来源于不同国家的科学文献产生引用关系的几率是相同国家的2.344($=e^{0.852}$)倍。模型拟合结果说明引用更倾向于在不同国家文献之间产生,这与研究假设预期不同。值得注意的是,虽然模型拟合结果否定了国家层面的文献同配效应,但是在前述描述性分析中也发现在少数国家中呈现出的同配趋势,例如表2、表3的分析都说明以美国为代表的几个国家具有非常高的引文吸引力,这些国家的学科文献不但内部互引率很高,而且也成为其他大部分学科水平相对落后国家的文献引用目标。在国家层面,科学研究存在着显著的马太效应[21],少数国家成为学科的核心与权威,引用从其他外围国家向核心国家聚集。个别高影响力国家成为引用网络的吸引奇点,不但吸引低影响力国家文献的引用,而且自身内部也存在高比例自引,引用沿着国家影响力分布的梯度方向流动。结合两方面分析,可以得出在整体上引用呈现出国家异配效应,而个别学科领先国家文献则呈现出一定程度的同配性。基于以上,实证结果部分支持假设5成立。

作为控制变量,*Nodeicov*(*CT*)的系数显著为正,文献学术水平每增加1个单位,其被引用的几率提高1.066倍。估计结果说明文献学术价值越高越有有助于增加新的被引关系,学术价值在引用关系形成过程中发挥正向影响,符合对于引文价值指向的基本认知与预期。

4 结论

本文针对施引文献与被引文献之间的相似性倾向,采用指数随机图模型,以图书馆与情报学领域为对象,开展实证研究。研究发现:①整体上,文献相似现象在引用关系中具有普遍性和多种表现形式,三角结构趋同、期刊同配、机构同配对引用关系具有促进效应;②深入分析发现文献相似倾向也表现出复杂性,例如在整个样本中,引用文献呈现出国家异配特征,但如果仅考虑占据学科优势国家的部分样本,则呈现出了文献间的国家同配倾向。

通过实证结果可以得到以下启示:

(1)社会接近性是引用关系的重要形成机制。相似文献表现出的引用偏好某种程度上也是社会接近性的体现。从信息搜索的角度看,相似的文献在搜索和辨别的机会成本上具有优势。面对海量的文献、复杂的学术问题,不论在哪个层面的社会接近性都为规避误判风险、避免学术偏差提供了高效率、低成本的方案指引,作者在时间、精力、学识的约束条件下,可能更愿意去相信和选择社会距离近的文献。从社会接近性的视角不但能够给文献相似现象给予较好解释,更丰富了引用形成机制的理解。

(2)伴随引用关系的文献相似是个体智力创造与群体社会因素交织的结果。引用本质上是一种特定的关系,是要素相互影响和交互的结果,例如观点的继承与碰撞、学术规范惯例的约束、作者间影响力展现、组织机构内强关系的延伸、地理空间的知识溢出效应、语言文化的兼容和惯性。各种要素群体交互形成的影响力不仅难以忽略,甚至可能超出了通常的预期,例如文献机构同配性所表现出的显著而强烈效应。虽然结论建立在特定学科样本之上,但亦充分反映出学术创作过程不单是逻辑、观点交互碰撞的思维活动,更是作者所嵌入社会网络环境的综合作用结果,而网络强关系产生的影响尤为重要。需要结合以科学社会学思想为代表的规范理论与社会建构理论加以解释。

(3)网络嵌入视角可能更有利于揭示引用行为的复杂性。引用行为具有复杂的动机,而引用关系之间存在的相关性某种程度上正是复杂性的体现,这一点

也通常被多数文献所忽视。引用关系在网络中涌现出结构嵌入特征体现了引用行为中作者群体的交互影响,实证研究也充分说明了引用关系倾向于嵌入三角结构的事实。社会网络和复杂网络理论为网络嵌入视角的分析提供了指导,而且随着大数据技术的日益成熟,建立在大数据基础上的网络建模会是未来深入揭示引用本质的有效途径。

因此,引文指标所测度的不仅仅是学术价值,更准确地讲是文献的综合影响力,是多种隐性要素影响的结果,非常有必要以客观谨慎的态度对待引文指标的适用和解释。另外,需要说明的是本文聚焦于社会层面的文献相似,未涉及文献主题或内容上的相似,可以扩展研究思路,将两个层面的相似性纳入同一研究框架,探讨知识流动过程中的社会建构、学术规范之间的交互关系及影响效果。后续研究有必要将实证扩展到其他学科领域,以检验研究结论的可靠性。

参考文献:

[1] 鲁索, 全薇. 期刊影响因子,旧金山宣言和莱顿宣言:评论和意见[J]. 图书情报知识, 2016(1):4 – 14.

[2] 王菲菲. 发文与引文融合视角下的科学计量学领域核心作者影响力分析[J]. 科学学与科学技术管理, 2014(12):45 – 55.

[3] COLIZZA V, FLAMMINI A, SERRANO M A, et al. Detecting rich-club ordering in complex networks[J]. Nature physics, 2006, 2(3):110 – 115.

[4] 金铁成. 学术期刊自引率使用乱象及其应对策略[J]. 科技与出版, 2016(11):96 – 98.

[5] ZHIVOTOVSKY L A, KRUTOVSKY K V. Self-citation can inflate h-index[J]. Scientometrics, 2008, 77(2):373 – 375.

[6] GLÄNZEL W, DEBACKERE K, THIJS B, et al. A concise review on the role of author self-citations in information science, bibliometrics and science policy[J]. Scientometrics, 2006, 67(2):263 – 277.

[7] BOOKSTEIN A, YITZHAKI M. Own-language preference: a new measure of "relative language self-citation"[J]. Scientometrics, 1999, 46(2):337 – 348.

[8] REN S, ROUSSEAU R. International visibility of Chinese scientific journals[J]. Scientometrics, 2002, 53(3):389 – 405.

[9] 唐莉, PHILIP S, YOUTIE J. 中国科研成果的引用增长是否存在"俱乐部效应"?[J]. 财经研究, 2016, 42(10):94 – 107.

[10] 马凤,武夷山. 关于论文引用动机的问卷调查研究——以中国期刊研究界和情报学界为例[J]. 情报杂志,2009(6):9 – 14.

[11] ROBINS G, SNIJDERS T, WANG P, et al. Recent developments in exponential random graph (p *) models for social networks[J]. Social networks, 2007, 29(2):192 – 215.

[12] PENG T Q. Assortative mixing, preferential attachment, and triadic closure: longitudinal study of tie-generative mechanisms in journal citation networks[J]. Journal of informetrics, 2015, 9(2): 250 – 262.

[13] CIMENLER O, REEVES K A, SKVORETZ J. An evaluation of collaborative research in a college of engineering[J]. Journal of informetrics, 2015, 9(3):577 – 590.

[14] BORNMANN L, DANIEL H. What do citation counts measure? a review of studies on citing behavior[J]. Journal of documentation, 2008, 64(1):45 – 80.

[15] ROUSSEAU R, LIU Y. Interestingness and the essence of citation [J]. Journal of documentation, 2013, 69(4):580 – 589.

[16] 王晰巍, 杨梦晴, 王楠阿雪, 等. "互联网 +"环境下美国iSchool院校科研项目发展动态研究[J]. 情报科学, 2017(3): 157 – 163.

[17] 向希尧, 裴云龙. 地理接近性对跨国专利合作的影响——社会接近性的中介作用研究[J]. 科学学与科学技术管理, 2016, 37(4):17 – 24.

[18] SNIJDERS T A B, PATTISON P E, ROBINS G L, et al. New specifications for exponential random graph models[J]. Sociological methodology, 2006, 36(1):99 – 153.

[19] 邱均平, 陈晓宇, 何文静. 科研人员论文引用动机及相互影响关系研究[J]. 图书情报工作, 2015,59(9):36 – 44.

[20] OPSAHL T, COLIZZA V, PANZARASA P, et al. Prominence and control: the weighted rich-club effect[J]. Physical review letters, 2008, 101(16):168702

[21] YANG X, GU X, WANG Y, et al. The Matthew effect in China's science: evidence from academicians of Chinese Academy of Sciences[J]. Scientometrics, 2015, 102(3):2089 – 2105.

作者贡献说明:

段庆锋:论文选题、数据分析、实证及论文撰写;
潘小换:数据采集及处理。

Empirical Research on Impact of Documents Similarity upon Scientific Citation Preference

Duan Qingfeng[1] Pan Xiaohuan[2]

[1] School of Management, Shanxi University of Finance & Economics, Taiyuan 030006

[2] School of Economic and Management, North University of China, Taiyuan 030051

Abstract: [**Purpose/significance**] It is possible to understand the mechanism of citation more deeply and clearly, understanding the nature of phenomenon in similarity between citing document and cited document which share the same or similar features. [**Method/process**] Aiming to discover the extent to which similarity between documents existing citation

relationship affects the preference of citations, empirical research was conducted to focus on the academic field of LIS by the statistical method of Exponential Random Graph Models (ERGM). [**Result/conclusion**] Some empirical results were found as following that there obviously exists tendency to be similarity between documents in the aspects of embedded network structure, affiliation and journal. Specially, the dyadic citation relation would be more likely to be embedded in the triangle transmit structure in citation network, and happened between the documents with the same affiliation and journal. Moreover, the documents, from the countries which are domain in the specific academic field, would be more likely to form the relationship of citation. Consequently, the empirical results adequately imply that social affinity is the crucial mechanism for citation behavior and reflect the social feature of citation preference.

Keywords: documents　similarity　scientific citation　ERGM

《图书情报工作》2017 年优秀审稿专家

2017 年,有 300 余位外审专家参加了《图书情报工作》稿件的同行评议工作,共评审稿件 2 000 余篇次,审阅 4 篇及以上的有 155 位,平均审稿时间为 6 天,高效、高质量的评审为《图书情报工作》遴选高质量稿件提供了保障。综合考虑今年以来的审稿数量、质量和时效,评选出 65 位优秀审稿专家(名单如下)。《图书情报工作》为优秀审稿专家颁发证书并免费寄送一年的期刊。感谢所有审稿专家对《图书情报工作》的大力支持!

(以下优秀审稿专家按姓氏拼音排序):

安小米　中国人民大学数据工程与知识工程教育部重点实验室
曹锦丹　吉林大学公共卫生学院
常　春　中国科学技术信息研究所
储节旺　安徽大学图书馆
邓胜利　武汉大学信息管理学院
丁　堃　大连理工大学 21 世纪发展研究中心暨 WISE 实验室
范爱红　清华大学图书馆
甘春梅　中山大学资讯管理学院
高　凡　西南交通大学图书馆
郭春侠　安徽大学管理学院
郭　宇　吉林大学管理学院
韩　毅　西南大学计算机与信息科学学院
何　胜　江苏理工学院计算机工程学院
胡昌平　武汉大学信息资源研究中心
胡正银　中国科学院成都文献情报中心
黄国彬　北京师范大学政府管理学院
黄　崑　北京师范大学政府管理学院
黄令贺　河北大学管理学院
姜春林　大连理工大学人文与社会科学学部
李　刚　南京大学信息管理学院
李国俊　北京大学图书馆
李　晶　安徽大学管理学院
李　睿　四川大学公共管理学院
李　武　上海交通大学媒体与设计学院
李月琳　南开大学商学院
刘　冰　天津师范大学管理学院
刘春丽　中国医科大学新校区图书馆
刘　华　上海大学图书馆
刘建准　天津工业大学管理学院
刘　勘　中南财经政法大学信息与安全工程学院
刘晓娟　北京师范大学政府管理学院
刘　宇　云南大学历史与档案学院
刘玉仙　同济大学图书馆
刘兹恒　北京大学信息管理系
牟冬梅　吉林大学公共卫生学院
裴　雷　南京大学信息管理学院
秦　鸿　电子科技大学图书馆
盛小平　华南师范大学经济与管理学院
宋　歌　东南大学图书馆情报科学技术研究所
苏新宁　南京大学信息管理学院
滕广青　东北师范大学计算机科学与信息技术学院
王翠萍　东北师范大学计算机科学与信息技术学院
王建芳　中国科学院科技战略咨询研究院
王立学　中国科学技术信息研究所
吴建华　华中师范大学信息管理学院
吴振新　中国科学院文献情报中心
吴志荣　上海师范大学图书馆
武夷山　中国科技发展战略研究院
谢　蓉　上海对外经贸大学图书馆
许海云　中国科学院成都文献情报中心
许　鑫　华东师范大学商学院
闫　慧　中国人民大学信息资源管理学院
杨建林　南京大学信息管理学院
杨思洛　武汉大学信息管理学院
俞立平　浙江工商大学管理工程与电子商务学院
袁顺波　嘉兴学院商学院
查先进　武汉大学信息管理学院
詹庆东　福州大学图书馆
张广钦　北京大学信息管理系
赵　飞　北京大学图书馆
赵宇翔　南京理工大学经济与管理学院
郑德俊　南京农业大学信息管理系
郑巧英　上海交通大学图书馆
周春雷　郑州大学信息管理学院
周庆山　北京大学信息管理系